Feature Selection

There are mainly two reasons for feature selection. Some features might be highly irrelevant or redundant. So it's better to remove these features to simplify the situation and improve performance. Additionally, it may seem like engineering more features is always a good thing, but as we mentioned earlier, many machine learning algorithms suffer from the *curse of dimensionality*—that is, they do not perform well when given a large number of variables or features.

We can improve the situation of having too many features through **dimensionality reduction**.

Commonly used techniques are:

- PCA (Principal Component Analysis)
- t-SNE (t-Distributed Stochastic Neighboring Entities)
- Feature embedding

Azure ML prebuilt modules:

- Filter-based feature selection: identify columns in the input dataset that have the greatest predictive power
- Permutation feature importance: determine the best features to use by computing the feature importance scores

QUIZ QUESTION

Below are the examples of dimensionality reduction algorithms that we just described. Can you match each one with its description?

Submit to check your answer choices!

DESCRIPTION	ALGORITHM
A linear dimensionality reduction technique based mostly on exact mathematical calculations.	PCA (Principal Component Analysis)
Encodes a larger number of features into a smaller number of "super-features."	Feature embedding
A dimensionality reduction technique based on a probabilistic approach; useful for the visualization of multidimensional data.	t-SNE (t-Distributed Stochastic Neighboring Entities)

SUBMIT

NEXT